Université Paris Diderot - Paris 7 2017-2018

Master 2 Biologie-Informatique/ Bioinformatique

Classification pour l'étude des séquences répétées centromériques chez les primates

Sarah Kaddah

Tuteur : Loïc Ponger

Structure et Instabilité des Génomes

MNHN - CNRS UMR 7196 / INSERM U1154 - Sorbonne Universités

Muséum national d'Histoire naturelle, 43 rue Cuvier 75005 PARIS

Remerciements

Je tiens tout d'abord à remercier énormément Loïc Ponger, responsable de mon stage, pour son encadrement, ses conseils, ses relectures et son aide.

Je tiens également à remercier Christophe Escudé pour tous ses conseils et pour la relecture de mon rapport, ainsi que l'équipe ARChE.

Je souhaite aussi remercier Evelyne Duvernois-Berthet, pour ses conseils tant au niveau professionnel que personnel, pour les discussions et pour les bonbons.

Je remercie aussi chaleureusement tout le laboratoire pour son accueil cordial.

Je remercie le journal club pour ces petites séances d'anglais sympathiques.

Je souhaite également remercier ici Catherine Etchebest et Jean-Christophe Gelly, et l'ensemble de l'équipe pédagogique, pour cette année de master 2.

Table des matières

1 Introduction

1.1 Le centromère

Le centromère est une structure chromatinienne essentielle au bon déroulement de la division cellulaire chez les eucaryotes. Il permet l'attachement du fuseau mitotique et la ségrégation des chromosomes [1]. Le kinétochore est un assemblage de protéines situé au niveau du centromère. Il va permettre l'attachement des microtubules aux chromosomes [2]. La chromatine centromérique est caractérisée par la présence de la protéine CENP-A, un variant de l'histone H3, très conservé au cours de l'évolution. Celle-ci fixe la position du kinetochore par un mécanisme épigénétique encore mal connu [3].

Bien que la fonction du centromère et des protéines sous-jacentes soit relativement bien conservée, les séquences d'ADN associées varient d'un taxon à l'autre. Cependant une structure commune se distingue étant de l'ADN répété en tandem, appelée ADN satellite [4]. Ces séquences peuvent représenter environ 5% du génome et la taille des unités de répétitions peut varier entre 7pb et 3,2kb [5].

1.2 L'ADN α-satellite

Chez les Primates, l'ADN satellite est connu sous le nom d'α-satellite. Ces séquences répétées en tandem sont riches en AT et un monomère fait 171 pb de long environ [6]. Ces séquences ont été mises en évidence pour la première fois chez *Chlorocebus aethiops* dans les années 1970 [7]. Des homologues ont été retrouvés chez d'autres espèces de primates [8]. Néanmoins ces séquences ont été essentiellement étudiées chez l'homme.

Les séquences ont un taux d'identité qui varie de 60 à 100% [9]. Les séquences les plus similaires peuvent être regroupées en familles (Figure 1). Ces familles résultent d'un même événement d'amplification. Des études chez l'homme ont montré que les séquences d'une même famille se regroupent phylogénétiquement mais aussi spatialement le long d'un chromosome [10]. Les familles se disposent de façon symétrique autour du cœur du centromère en respectant un gradient d'âge. Ces observations ont permis de proposer un modèle évolutif avec des centromères en expansion. Les familles les plus récentes s'insèreraient au cœur du centromère, dans la partie active, repoussant les familles les plus anciennes jusqu'aux régions voisines, appelées péri-centromères. Les familles les plus anciennes ont une organisation monomérique, sous forme de longs segments composés de monomères de la même famille, tandis que les fa-

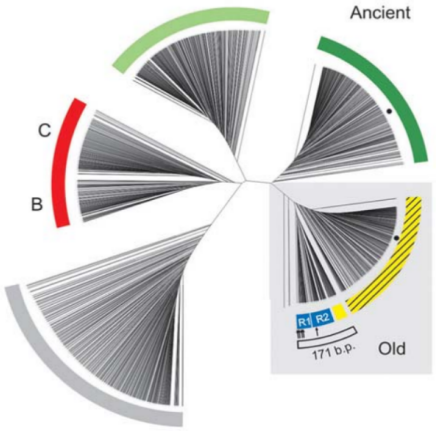

FIGURE 1 – **Arbre phylogénétique des séquences α-satellites du bras p du chromosome X :** L'arbre réunit 1431 monomères présents sur le contig nt011630 [10].

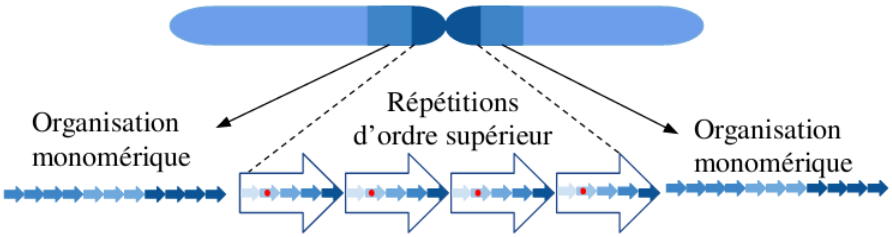

FIGURE 2 – **Organisation spatiale des α-satellites :** Le cœur du centromère (bleu foncé) est organisé en répétition d'ordre supérieur. Le péricentromère (bleu clair) a une organisation monomérique. Un monomère d'une même famille est représenté par une petite flèche de même couleur. Les points rouges représentent les sites de fixation à CENP-B ou pJα.

milles récentes sont organisées en structure d'ordre supérieur ou *Higher Order Repeat* (HOR), où un groupe de monomères appartenant à des familles différentes sont répétées en bloc les un derrière les autres (Figure 2).

Le rôle des α-satellites est encore mal connu dans la fonction du centromère. Seule la protéine CENP-B serait capable de reconnaître spécifiquement un motif d'environ 17 pb présente sur certains monomères (CENP-B box). Bien que cette protéine soit présente chez tous les organismes, la CENP-B box est remplacée par un autre motif dans certaines familles. Ce motif serait capable de lier une protéine très mal caractérisée nommée pJα (pJα box) [11]. Il arrive que ces motifs soient remplacés par le motif pKβ (pJβ box).

1.3 Le sujet de stage

L'objectif de ce stage est d'étudier les α-satellites à partir de données de séquençage haut débit afin de comprendre la fonction de ces séquences.

Peu d'informations sur les α-satellites existent chez les autres espèces de primates, et aucune relation inter-espèce n'a été réalisée, le séquençage et l'assemblage de ces séquences étant difficiles. Jusqu'à présent, les études se basant sur un séquençage haut-débit ont été appliquées chez l'homme (cité ci-dessus) et chez le Gorille. L'équipe d'accueil de mon stage "ADN répété, Chromatine, Evolution" ou ARChE, a récemment développé une approche de séquençage haut débit, ciblée sur les séquences α-satellites chez le *Cercopithecus solatus* et le *Cercopithecus pogonias*. Ces deux espèces ont beaucoup d'ADN satellite et de réarrangements chromosomiques, avec l'apparition de nombreux centromères. Cette étude a également montré les limites des approches classiques (alignements et phylogénie). Ces méthodes ne permettent pas de traiter des jeux de données de grande taille, or il peut y avoir plusieurs milliers de monomères dans un seul génome. De plus, ces méthodes non-objectives ne permettent pas de faire des comparaisons entre espèces.

Pour remédier à ce problème, une méthode de classification automatisée des α-satellites a été développée dans l'équipe. Ce programme permet de traiter des centaines de milliers de séquences, quelque soit le nombre ou la taille des familles. De plus, cette méthode est objective et peut être appliquée à plusieurs espèces, permettant ainsi une comparaison inter-espèce des familles α-satellites.

Mon sujet consiste à appliquer cette méthode aux jeux de données déjà publiés pour évaluer la méthode. Dans une deuxième temps, cette méthode est appliquée à deux autres primates dans le but de caractériser les familles d'espèces proches. Les mécanismes d'évolution pourront être déduits à partir d'une comparaison inter-espèce révélant les différences et les familles communes.

2 Matériel et méthode

2.1 Les espèces étudiées

Les données de *c. solatus* et *C. pogonias* sont issues d'un séquençage ciblé de monomères digérés par une enzyme de restriction (XmnI, HindIII) et séquencés en Ion Torrent [13]. Les reads sont issus du séquençage à haut débit, dont Les données sont disponibles dans Genbank (septembre 2017). Les monomères d'α-satellites faisant environ 171pb, seuls les projets impliquant des reads relativement longs ont été considérés (illumina en single end, L454, ...). Les critères de sélection dépendent de la disponibilité des séquences de qualité parmi 10 espèces.

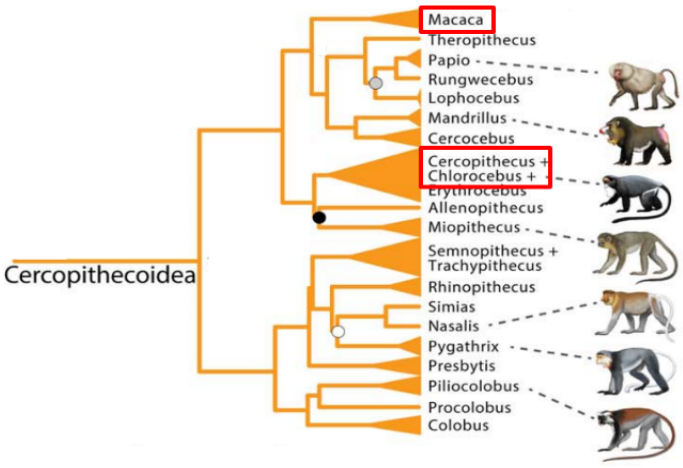

FIGURE 3 – **Arbre phylogénétique des espèces analysées.**[12]

Afin de permettre de comparer les données avec les deux cercopithèques, j'ai choisi deux espèces proches parmi les 10 espèces pour lesquelles des données sont disponibles,(Figure 3), le *Macaca fascicularis* et le *Chlorocebus sabaeus*. Tous les α-satellites sont alignés sur les monomères de *C. solatus* et *C. pogonias* étudiés précédemment.

2.2 Méthode de classification

2.2.1 Principe

Cette méthode a été développée par Florence Jornod (stage M2, 2016-2017). Elle répartit des séquences α-satellites en familles selon la similarité calculée à partir de leur composition en 5-mers. La classification est hiérarchique dichotomique. Au départ, une table contenant les fréquences des 5-mers est calculée pour chaque monomère. Ensuite une boucle itérative est exécutée pour séparer les séquences en deux groupes. Cette étape est répétée tant que les nouveaux groupes formés sont divisibles.

2.2.2 Répartition itérative

Une Analyse en Composante Principale (ACP) est effectuée sur la table des fréquences des 5-mers afin de réduire les dimensions du jeu de données et d'obtenir des variables indépendantes. Des distances euclidiennes sont calculées entre toutes les paires de séquence dans l'espace défini par les premières composantes de l'ACP.

A partir du calcul de distance, les séquences sont séparées en deux classes en utilisant la classification hiérarchique basée sur la méthode de Ward. Cette méthode maximise l'inertie in-

terclasse. La classification hiérarchique fait un usage important de la mémoire. Par conséquent, pour traiter des jeux de données importants. La classification est appliquée à 100 000 séquences choisies aléatoirement et une analyse discriminante linéaire (LDA) est utilisée pour classer les autres les séquences.

2.2.3 Double-validation d'un sous-groupe

A chaque itération, la classification hiérarchique divise le jeu de données en deux groupes qui sont évalués avant d'être ou non redivisés. Afin d'éviter de diviser les données en familles trop petites, le premier critère de validation est la taille du sous-groupe. Si un groupe atteint 100 séquences, il n'est pas redivisé. Le deuxième critère de validation s'appuie sur le *matepair* et permet d'estimer la qualité de la partition. Ce terme correspond à la proportion de monomères ayant son plus proche voisin dans la même classe, se basant sur les distances euclidiennes calculées auparavant. Ainsi des valeurs de *matepairs* élevées (proches de 1) indiquent des sous-groupes bien homogènes et séparés validant la classification.

Le seuil de *matepair* est fixé à 0.90, pour avoir des groupes homogènes. Si au moins une des valeurs de *matepair* est au-dessous de ce seuil, les sous-groupes sont considérés comme formant un seul groupe et le groupe initial est sauvegardé comme une famille unique. Si les *matepairs* sont au-dessus d'un certain seuil, les deux sous-groupes sont ajoutés séparément à la file pour être potentiellement redivisés ultérieurement.

2.3 Analyse des séquences

Les séquences monomériques sont comparées à partir de leur composition en 5-mers dans le but d'identifier des regroupements d'α-satellites sans passer par l'étape d'alignement. Pour chaque ensemble de monomères, la table de fréquence des 5-mers est analysé en utilisant l'Analyse en composante principale pour réduire l'espace de complexité pour pourvoir visualiser les données sur les premiers plans factoriels.

L'alignement des séquences est fait avec Muscle [14] et visualisé avec Seaview (problème de biblio UTF8). La phylogénie est construite avec la méthode du maximum de vraisemblance (PhyML) [**?**]. Le modèle F84 est utilisé pour la construction de l'arbre. Le support de branche est aLRT (SH-like). La fréquence d'équilibre des nucléotides, le ratio de transition et de transversion et les taux de variation sont optimisés.

Les consensus sont obtenus avec des scripts développés par l'équipe. Les motifs CENP-B

TABLE 1 – Résumé du jeu de données et des résultats de la classification.

Espèce	Nb séq tot	Nb fam tot	Nb fam > 100 séq	% seq ds fam > 100 seq
C. solatus	105 529	564	12	96.03
C. pogonias	112 902	132	13	98.71
C. sabaeus	29 842	338	43	89.11
M. fascicularis	235 535	3694	114	88.94

(TTCGTTGGAA[AG]CGGGA), PJα (TTCCTTTT[CT]CACC[AG]TAG) et pKβ (GATATCCCGGTTTCCTT) ont été identifiés avec le logiciel fuzznuc (package EMBOSS) [15] et en autorisant 2 différences au maximum par rapport au consensus.

3 Résultats

3.1 Caractérisation intraspécifique des familles

Les séquences ont été classées en utilisant une méthode de classification objective développée dans l'équipe. Le nombre de familles est déterminé indirectement par un critère sélectionnant des classes à la fois homogènes et différentes les unes des autres.

3.1.1 Identification des familles

A l'issue de la classification, seules les familles ayant plus de 100 séquences, appelées grandes familles, sont conservées pour l'analyse. Pour les 4 espèces, le nombre de familles conservées diminue considérablement après élimination des petites familles (<100 séquences) mais ces familles ne représentent que 11% des séquences du jeu de données au plus (Tableau 1). Malgré le nombre de familles qui diffère d'une espèce à l'autre, la distribution des familles est similaire chez les quatre espèces. Les plus petites familles sont très nombreuses et et la fréquence diminue quand la taille augmente (Figure 4).

Les espèces *C. solatus* et *C. pogonias* sont analysées dans un premier temps pour comparer la classification automatisée avec la classification empirique faite dans l'équipe. Ces familles ont été définies manuellement à partir d'une ACP basée sur la composition en 5-mers de tous les monomères (Fig. 5). 6 familles α-satellites ont été définies empiriquement et confirmé expérimentalement chez les Cercopithèques. Ces deux espèces partagent deux grandes familles monomériques, nommées C1 et C2, et deux familles formant un HOR d'ordre 2, nommées C3-C4, de l'ordre d'une centaine de séquences chacune. *C. pogonias* possède les familles supplé-

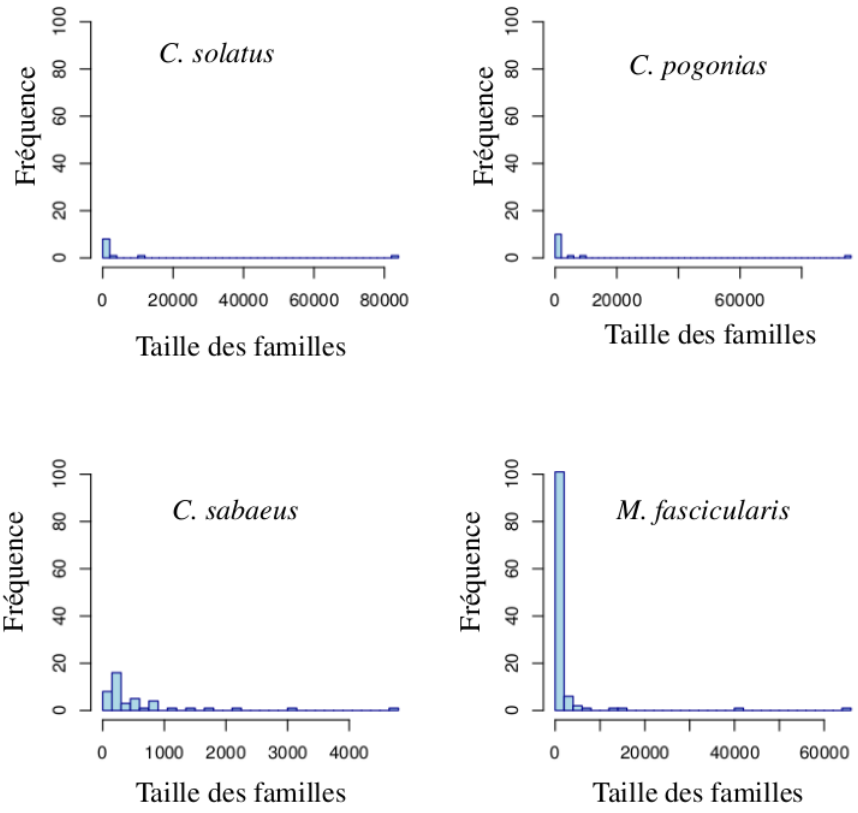

FIGURE 4 – **Distribution des familles en fonction du nombre de séquences.**

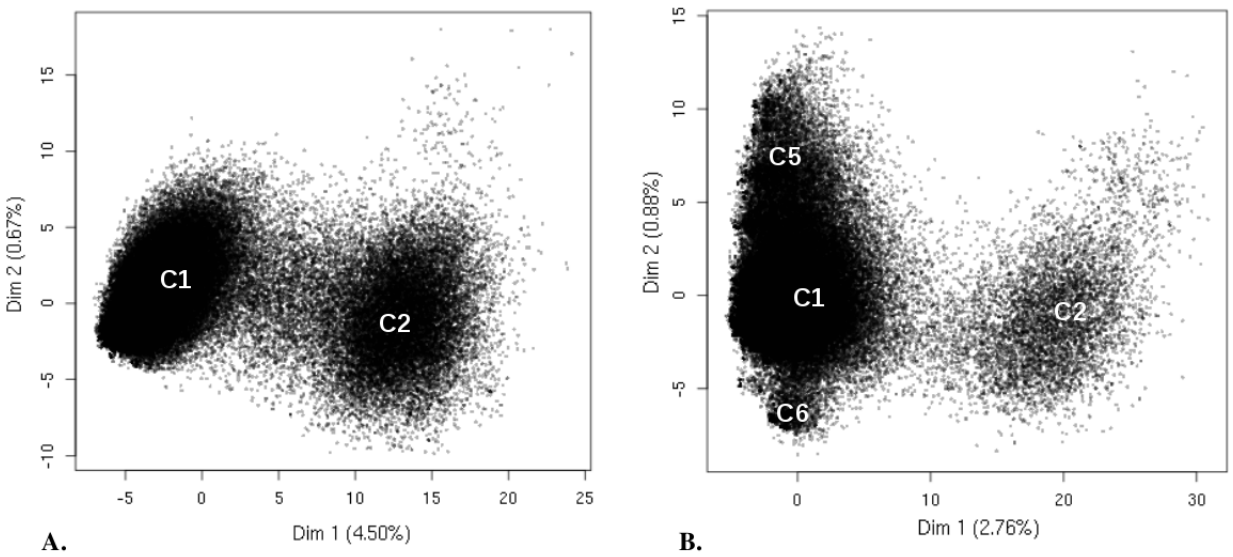

FIGURE 5 – **Caractérisation visuelle des familles α-satellite chez *C. solatus* et *C. pogonias* à partir d'une ACP, basée sur la composition en 5-mers des monomères :**Le nom des familles est indiqué sur les graphiques. Un point représente un monomère. **A.** *C. solatus.* **B.** *C. pogonias.*

Classification publiée	*C. solatus*	*C. pogonias*
C1	1	10
C2	11	1
C3	1*	1
C4	1*	1*
C5	-	1
C6	-	0
Total	12 + 2* < 100 seq	13 + 1* <100 seq

TABLE 2 – **Comparaison des classifications automatiques avec celles publiées précédemment (Cacheux et al, 2016 et 2018)**

mentaires C5 et C6.

Bien que le nombre de grandes familles soit relativement proche entre ces deux espèces, les résultats diffèrent significativement (Tableau 2). Chez *C. solatus* les familles C1 à C4 sont retrouvées : 11 familles appartiennent à C2, une famille appartient à C1 et les familles C3 et C4, ne comportant que 109 et 112 séquences dans le jeu de données, sont retrouvées dans des petites familles d'environ 80 séquences chacune. Chez *C. pogonias*, contrairement à C6, les familles C1, C2, C3, C4 et C5 sont retrouvées. C1 est répartie en 10 familles, alors que les séquences de C2, C3 et C5 regroupées dans des familles individuelles, et la famille C4 est également retrouvée sous la forme d'une petite famille de 86 séquences. Chez le *C. solatus*, la famille C2 est divisée en plusieurs familles et la famille C1 est retrouvée dans une seule famille. La situation inverse est retrouvée chez *C. pogonias*.

La diversité de ces familles a été analysée avec une ACP basée sur sur la composition en 5-mers (Figure 6). Afin de valider la surclusterisation observée pour les séquences de C2 chez *C. solatus*. La pertinence de la classification est confirmée par l'ACP ou les phylogénies. Chez *C. solatus*, les séquences de C1 est entièrement retrouvée dans une famille. Le groupe C2 est réparti en plusieurs familles. Deux familles intermédiaires (rouge et bleue) sont visibles entre la famille C1 (vert) et C2 (orange). Elles ne sont pas distinctes. Une famille supplémentaire (turquoise) se démarque. Pour confirmer cette division des séquences de C2, la visualisation de l'ACP des 5-mers est observée en fonction des composantes 1 et 3. Les familles intermédiaires sont toujours confondues, contrairement à la famille turquoise qui forme une famille à part entière. Pour confirmer ce résultat, l'arbre construit atteste que chaque famille est bien retrouvée, notamment les familles intermédiaires qui forment bien deux familles. Chez *C. pogonias*, les familles C2, C4 et C5 sont bien retrouvées. La famille C6 se fond dans la famille C1 (en vert). La famille C1 est divisée en deux familles supplémentaires (rose et violet). La visualisation des

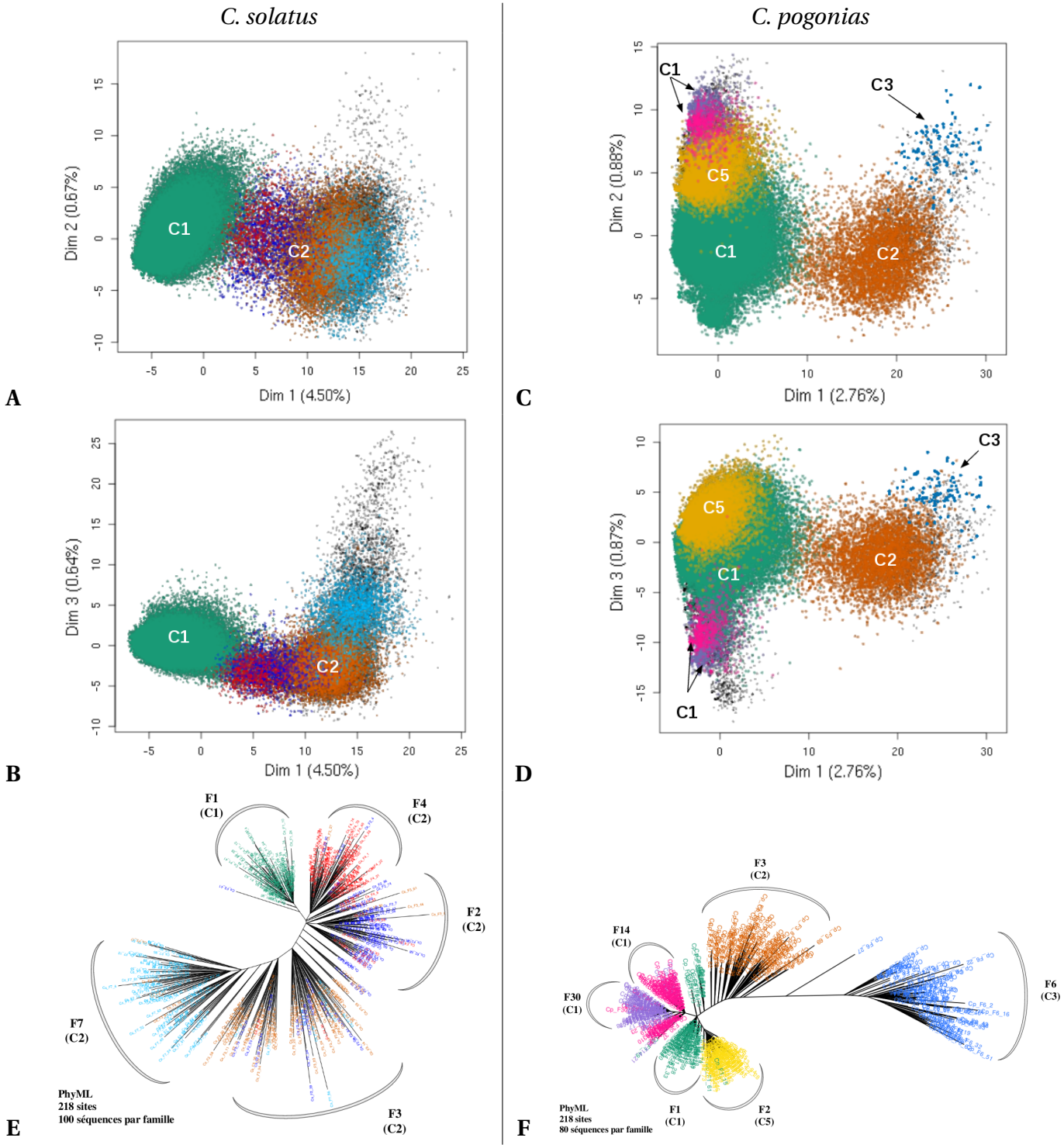

FIGURE 6 – **Représentation des plus grandes familles issues de la classification automatisée :** Ces familles sont superposées sur les représentations de l'ACP des 5-mers. **A.** Composantes 1 et 2 de l'ACP. Les familles qui correspondraient à C1 sont en vert, C2 en orange, rouge, bleu et turquoise. **B.** Composantes 1 et 3 de l'ACP. **C.** Composantes 1 et 2 de l'ACP. Les familles qui correspondraient à C1 sont en vert, violet et rose, C2 en orange, C4 en bleu clair et C5 en jaune. **C.** Composantes 1 et 3 de l'ACP. **E. et F.** Phylogénie des différentes familles chez *C. solatus* (100 séquences par famille) et *C. pogonias* (80 séquences par famille) respectivement. Les couleurs sont respectivement conservées.

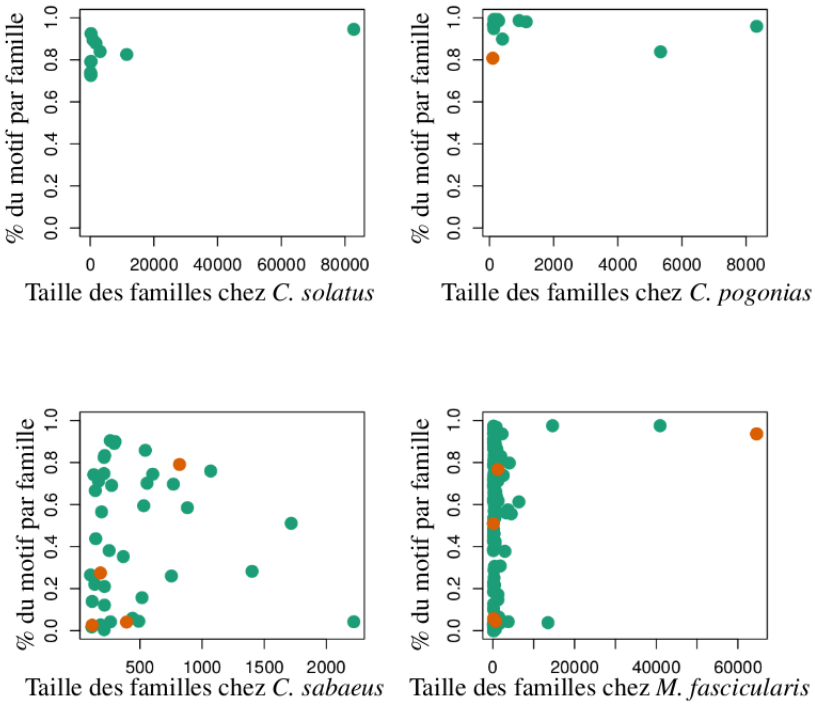

FIGURE 7 – **Fréquences des motifs CENP-B, pJα ou pKβ dans les familles :** Les motifs ont été cherchés sur la base de leur consensus en autorisant deux différences. Le pourcentage de séquences par famille ayant le motif pJα est en vert, pKβ en orange et CENP-B est absent de toutes les familles. Chaque famille est également représentée en fonction de sa taille.

composantes 1 et 3 de l'ACP ne permet pas de trancher sur la classification. L'arbre montre que les familles en rose et violet sont très proches.

En conclusion, la classification automatique permet de retrouver les familles identifiées dans les travaux publiés, excepté les séquences de C6 qui semblent indissociables de certaines séquences C1, et la mise en évidence de plusieurs familles parmi les C2 chez *C. solatus* et *C. pogonias*.

3.1.2 Motifs potentiellement fonctionnels

Une des fonctions des α-satellites pourrait être associée à fixer certaines protéines. A ce jour, deux motifs, CENP-B et pJα, mutuellement exclusifs et capables de fixer deux protéines ont été décrits. Des travaux en cours dans l'équipe suggèrent l'existence d'un troisième motif conservé, nommé pKβ, qui serait présent dans certaines familles. Ces trois motifs ont été recherchés sur la base de leur consensus (avec 2 différences autorisées).

La protéine CENP-B est présente chez toutes les espèces de primates, mais les Cercopithèques ne possèdent pas son site de liaison.En effet, CENP-B est absent chez *C. solatus* et *C. pogonias*. *C. sabaeus* et *M. fascicularis* n'ont pas ce motif non plus. La protéine pJα est une pro-

téine peu connue mais dont le site de liaison est déterminé. Le motif pKβ est un site commun aux familles n'ayant ni CENP-B ni pJα. Ces trois motifs sont recherchés pour chaque famille de chaque espèce sur la base des consensus (cf matériel et méthodes). Par contre plus de 90% des familles chez les quatre espèces ont le motif pJα mais à des niveaux différents.

La totalité des familles chez *C. solatus* ont le motif pα, dont 8 à plus de 75%. Chez *C. pogonias*, 12 familles sur 13 ont le motif à plus de 75%. Les séquences de C1 chez ces deux espèces ont le motif à 95%. *C. sabaeus* présente 39 familles avec ce motif, dont 7 l'ayant à plus de 75%. *M. fascicularis* a des pourcentages pour le motif pJα qui varie entre 1% et 97%, dont 28 familles avec le motif à plus de 75%.

Le motif pKβ est absent chez *C. solatus*. Une seule famille, incluant les séquences de C3 a ce motif à 80% chez *C. pogonias*. *C. sabaeus* a quatre familles avec le motif, dont deux à 79% et l'autre à 29%. *M. fascicularis* a cinq familles avec ce motif. La famille ayant le motif à 93% est une grande famille de 64 000 séquences. Les deux autres familles ont le motif à 0.76% et 0.50%.

Les quatre espèces ont des points en commun concernant l'absence du motif CENP-B et quelques familles ayant pKβ. Cependant pour le motif pJα, *C. solatus* et *C. pogonias* ont des pourcentages relativement proches mais qui diffèrent du *M. fascicularis* et du *C. sabaeus*, dont les valeurs sont intermédiaires. Chaque famille a un motif pour toutes les espèces excepté *M. fascicularis* qui a 6 familles sans motifs identifiables.

Pour conclure, le motif CENP-B est absent, le motif pJα est le plus présent le motif le mieux conservé, et le motif pKβ est rare mais présent dans quelques familles seulement, et absent chez *C. solatus*.

3.1.3 Similarité entre familles

Selon les hypothèses actuelles, les familles α-satellites seraient le résultat d'amplifications qui généreraient plusieurs centaines ou milliers de copies identiques au cours du temps. La divergence intra-famille a été estimée à partir d'un échantillon de 500 séquences par famille (au plus) choisies aléatoirement afin d'estimer l'âge des familles.

C. solatus et *C. pogonias* ont des pourcentages relativement faibles ne dépassant pas 15%, le *M. fascicularis* a des valeurs intermédiaires variant de 0.1% à 28%, et *C. sabaeus* a les valeurs les plus élevées allant de 11% à 28%.

Chez *C. solatus*, toutes familles ont entre 10% et 15% de divergence sauf celles comprenant les séquence de C1 qui au dessous de 6%. Concernant *C. pgonias*, les séquences de la famille

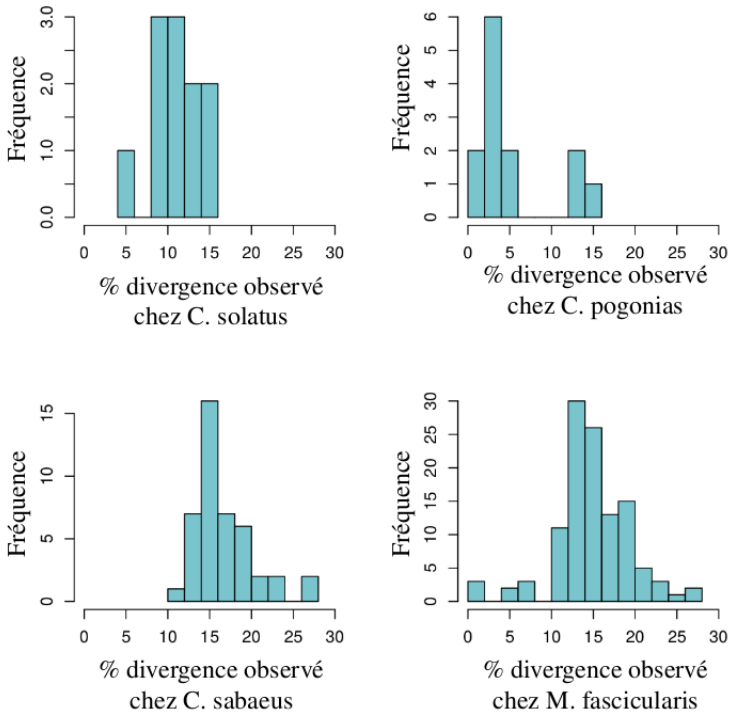

FIGURE 8 – **Pourcentage de divergence observée de chaque famille :** Histogramme du nombre de famille en fonction du pourcentage de divergence observé.

C2 et C3 sont au dessus de 7%. *C. sabaeus* n'a que des familles divergentes, le pourcentage de divergence étant supérieur à 10%. *M. fascicularis* a quelques familles avec un faible pourcentage de divergence, quelques familles autour de 15% et quelques unes qui se rapprochent de 30 %.

Pour conclure, une distribution bimodale est plus ou moins observée, avec des familles peu divergentes et d'autres plus divergentes.

3.2 Comparaison inter-espèce

Pour comparer les classifications obtenues au sein de chaque espèce, une super-classification a été effectuée afin de déterminer l'existence de super-familles (SF). Pour cela, l'algorithme de classification a été utilisé sur un jeu de donné composé de 100 séquences par grande famille (> 100 séquences) tirées aléatoirement pour chaque espèce. Un jeu de données de 18 100 séquences est soumis à la classification automatique. A l'issue de cette super-classification, 158 familles sont obtenues au total, dont 90 grandes familles (Figure 9). Les petites familles de moins de 20 séquences, soit 1.76% de ce jeu de données, ne sont pas prises en compte dans l'analyse. Après avoir présenté quelques caractéristiques de cette analyse, je présenterai les résultats concernant deux questions : l'origine du site de fixation de pKβ et la divergence des séquences appartenant à C2.

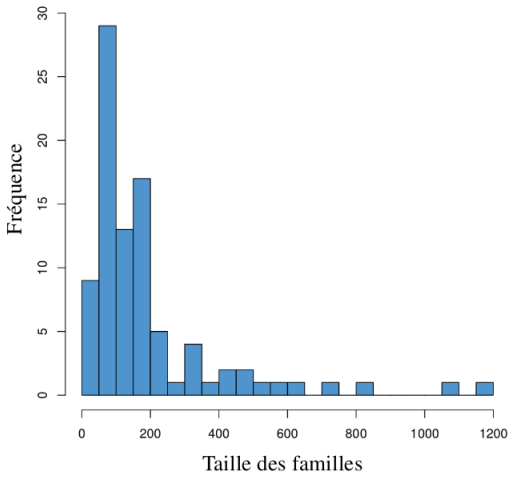

FIGURE 9 – **Distribution des super-familles :** Histogramme du nombre de famille observé en fonction de la taille.

3.2.1 Répartition des super-familles

Parmi les super-familles, seulement une *a priori* est commune aux quatre espèces. Elle rassemble 6 familles parmi les 10 familles classées C2 de *C. solatus* et les deux familles également classées C2 de *C. pogonias*, ainsi que deux familles du *M. fascicularis* et une famille du *C. sabaeus*. Une partie de la famille annotée C2 serait donc commune aux quatre espèces.

Une super-famille est partagée entre le *C. pogonias*, le *C. sabaeus* et le *M. fascicularis*. Cette super-famille regroupe des familles ayant le motif pKβ et qui correspondrait à la famille C3 commune au *C. pogonias* et au *C. solatus*. Cette famille de *C. solatus* est absente car elle ne fait que 80 séquences. Etant une famille de moins de 100 séquence, elle n'a pas été retenue pour l'analyse. Cela signifierait que la famille C3 serait commune à ces quatre espèces également.

Certaines super-familles sont spécifiques aux espèces. Au sein des 8 super-familles spécifiques du *C. pogonias*, seule l'une d'entre elles regroupe deux familles, le reste étant composé d'une famille. Elles correspondraient aux familles C5, une petite partie de C6 et essentiellement à C1. Les trois super-familles spécifiques de *C. solatus* seraient équivalentes aux familles C2. Le *C. sabaeus* en a 3 et le *M. fascicularis* 27.

Le *C. sabaeus* et le *M. fascicularis* partagent 38 super-familles, soit 42% des grandes super-familles (> 20 séquences), dont la plus grande faisant une taille de 1194 séquences. Une seule super-famille est uniquement commune à *C. solatus* et *C. pogonias* et elle regroupe les deux plus grandes familles qui seraient du C1.

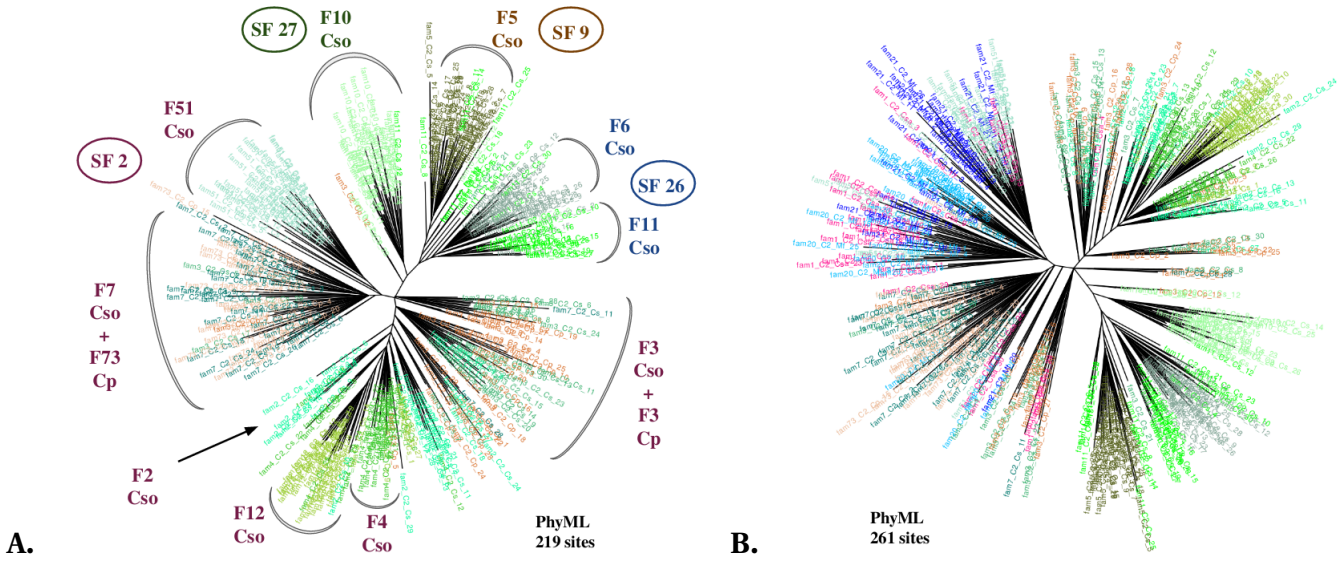

FIGURE 10 – **Phylogénie des différentes familles qui correspondraient à la famille C2 : A.***C. solatus* (tons verts) et *C. pogonias* (tons orange) **B.** Des familles du *M. fascicularis* (tons bleus) et *C. sabeus* (rose) sont rajoutées.

3.2.2 Un groupe C2 hétérogène

La famille C2 fait l'objet d'une séparation en plusieurs familles intéressantes. Pour vérifier les résultats de cette super-classification, 30 séquences de chacune des familles annotées C2 de *C. solatus* et *C. pogonias* sont tirées aléatoirement et un arbre est construit pour voir si cette division est retrouvée. Un autre arbre, avec les familles des deux autres espèces supposées de la famille C2, est construit.

Une partie de ces familles est spécifique au *C. solatus*, tandis que les familles restantes sont communes aux quatre espèces selon la classification. Dans la phylogénie, les super-familles 27, 9 et 26 sont bien retrouvées. Cependant la super-famille 2 regroupe trop de familles, sans faire de distinction (Figure 10 A). Les familles présumées C2 des deux autres espèces se mélangent plus spécifiquement avec la famille 51 de *C. solatus*.

3.2.3 Origine du motif pKβ

Pour poursuivre l'étude sur les familles communes entre espèces, le regroupement des familles α-satellites ayant le motif pKβ ou "familles pKβ" attirent l'attention.

La première super-famille pKβ est constituée de la famille C3 de *C. pogonias*, la famille 3 de *C. sabeus*, et la famille 8 de *M. fascicularis*. La super-famille 21 est spécifique au macaque, regroupant la famille 1. Toutes ces espèces ont donc une "famille C3" en commun, peut-être sous forme de dimère C2-C3.

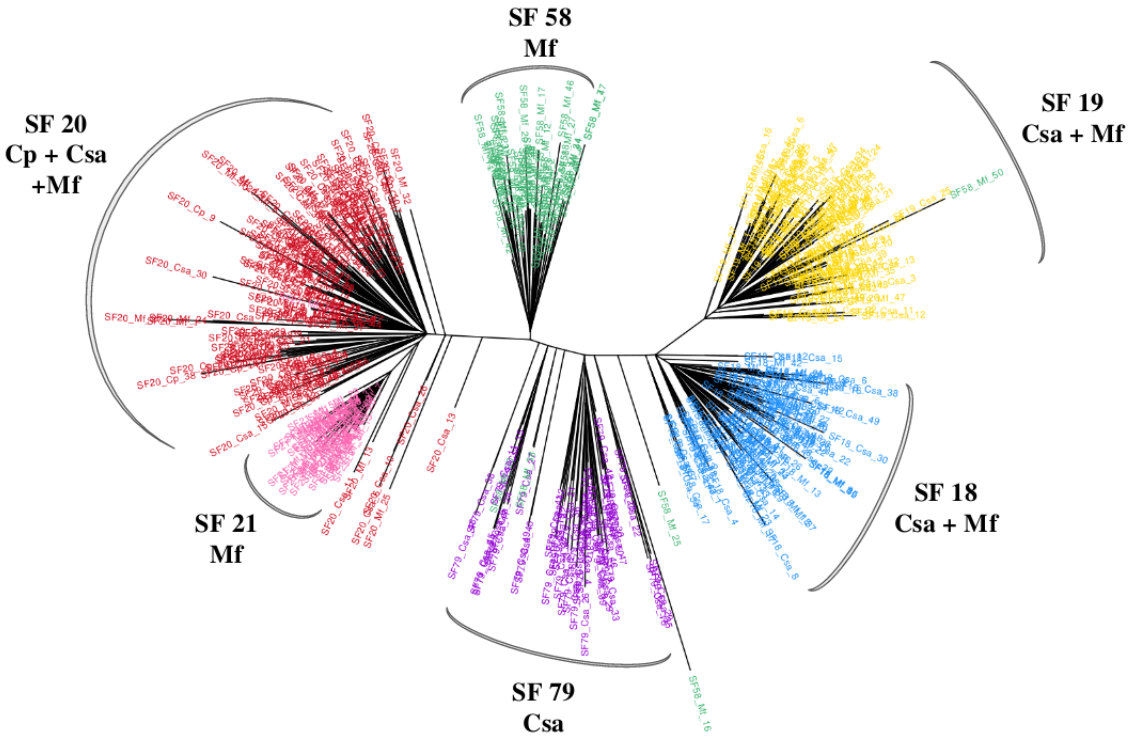

FIGURE 11 – **Phylogénie des super-familles pKβ**

Le *M. fascicularis* et *C.sabeus* ont deux super-familles pKβ en commun. Les familles constituant ces super-familles ont un pourcentage du motif relativement faible, pourtant elles sont tout de même rassemblées.

Le *C.sabeus* et le *M. fascicularis* ont respectivement chacun une super-famille pKβ spécifique. La première super-famille a le motif à 27% et la deuxième à 50%.

Un arbre composé de toutes les super-familles pKβ est construit pour voir comment elles s'assemblent. Le jeu de données est construit à partir de 50 séquences par famille pKβ par espèce tirées aléatoirement (Figure 11). Elles se rassemblent exactement comme dans les super-familles. La méthode de classification retrouve bien les familles pKβ et elle les classe correctement.

Pour conclure, les familles pKβ sont des familles partagées entre espèces pour certaines, et spécifique à une espèce pour d'autres.

4 Discussion

Après classification, les séquences α-satellites sont réparties en familles. Le nombre de familles par espèce est très variable. *C. solatus* et *C. pogonias* ont seulement un dizaine de familles

comparé au *M. fascicularis* qui en a un peu plus d'une centaine. *C. sabaeus* a un nombre intermédiaire de famille.

La méthode de classification crée beaucoup de petites familles (< 100 séquences). Par conséquent, plusieurs séquences sont éliminées et celles-ci pourraient être une famille caractéristique. Ces familles de petite taille sont probablement des séquences ayant subi des réarrangements, ayant une délétion importante, ou dont la qualité est moins bonne. De plus très peu d'information est perdue.

Bien que la méthode de classification ait retrouvé la majorité des familles empiriques, elle n'a pas réussi à dissocier les séquences C6 de C1. Par contre, elle a pu diviser C1 chez *C. pogonias* et C2 chez *C. solatus*. Cet algorithme de classification a du mal à séparer des séquences qui se seraient trop proches mais il peut séparer un groupe qui n'est pas parfaitement homogène. En effet les divisions du groupe C2 ont bien été retrouvées sur un arbre phylogénétique pour *C. solatus*.

La fonction la mieux conservée est celle derrière le motif pJα. Il est présent dans pratiquement toutes les familles, même si sa présence est moindre dans quelques une des familles chez *C. sabeus* et *M. fascicularis*. La fonction de pKβ, bien que largement moins présente, reste très conservée. De plus, certaines super-familles sont communes à plusieurs espèces. Parmi celles-ci, les plus anciennes sont la super-famille spécifique à *C. sabaeus* et celle spécifique au *M. fascicularis*, avec un pourcentage de divergence de 17,50% et 16,12% respectivement. La famille la plus jeune est la deuxième super-famille spécifique au macaque avec 7,09% de divergence. La super-famille commune aux 4 espèces est relativement ancienne, avec en moyenne 14,70% de divergence, et les familles communes à la fois au *M. fascicularis* et au *C. sabaeus* sont un peu plus jeune avec respectivement 12,50% et 11.80% de divergence.

Chez *C. solatus* et *C. pogonias*, les séquences C1 ont le pourcentage de divergence le plus faible, elles sont donc des familles jeunes. Les séquences de la famille C2, et C3 chez *C. poginias*, ont les pourcentages de divergence les plus élevés. Ce sont les familles les plus anciennes. La distribution bimodale pourrait supposer des vagues d'amplification.

La méthode de classification, confirmée par l'analyse phylogénétique, rassemble une famille de *C. sabaeus* et deux familles du *M. fascicularis* avec des séquences C2 de *C. solatus* et *C. pogonias*. Ces espèces ont donc des séquences qui seraient potentiellement des séquences C2. Les taux de divergence relativement élevés, autour de 14% pour ces familles, s'accorde avec cette information. Cependant il pourrait y avoir d'autres séquences de C2 qui ne seraient pas

détectées. Certaines de ces séquences forment une super-famille commune aux 4 espèces tandis que d'autres sont spécifiques à une seule espèce. Ces informations permettent de déduire l'ordre d'apparition des familles. La super-famille C2 commune aux 4 espèces serait une famille ancienne qui serait apparue bien avant les autres. Les familles un peu plus récentes seraient celles partagées entre *M. fascicularis* et *C. sabaeus*, et celles qui sont spécifiques aux espèces seraient les plus récentes. Bien que le *C. sabaeus* soit plus proche des cercipithèques, il partage plus de familles avec *M. fascicularis*. Il est possible que certaines familles de *C. solatus* et *C. pogonias* ne figurent pas parmi les résultats, car toutes les familles n'ont pas été détectées par l'enzyme de restriction.

Pour évaluer la reproductibilité, plusieurs super-classifications ont été lancées et analysées. Avec les mêmes paramètres, d'une super-classification à l'autre, les super-familles obtenues sont différentes. Seules les familles pKβ se rassemblent exactement de la même manière. Des séquences C2 sont regroupées, mais pas de la même manière. Elles peuvent former une seule famille et parfois plusieurs familles. Cette méthode de classification possède donc un problème de reproductibilité.

5 Conclusion

Pour conclure, les séquences α-satellites ont été réparties en familles, caractérisée en fonction de ce qui a été trouvé dans la littérature et en par la fonction des motifs caractéristiques, ainsi que par le pourcentage de similarité. La super-classification a permis de trouver les familles communes entre espèces et de déduire les familles les plus anciennes et les plus récentes ou encore celles qui sont spécifiques.

Il serait intéressant d'étudier la structure en HOR ou en organisation monomérique. Si les familles sont définies en une structure particulière, la question se pose de savoir si elles conservent la même structure d'une espèce à l'autre.

Références

[1] Don W Cleveland, Yinghui Mao, and Kevin F Sullivan. Centromeres and kinetochores : from epigenetics to mitotic checkpoint signaling. *Cell*, 112(4) :407–421, 2003.

[2] Stefano Santaguida and Andrea Musacchio. The life and miracles of kinetochores. *The EMBO journal*, 28(17) :2511–2531, 2009.

[3] Kevin F Sullivan, Mirko Hechenberger, and Khaled Masri. Human cenp-a contains a histone h3 related histone fold domain that is required for targeting to the centromere. *The Journal of cell biology*, 127(3) :581–592, 1994.

[4] S Henikoff, K Ahmad, and H S Malik. The centromere paradox : stable inheritance with rapidly evolving dna. *Science (New York, N.Y.)*, 293 :1098–1102, August 2001.

[5] A Cellamare, C R Catacchio, C Alkan, G Giannuzzi, F Antonacci, M F Cardone, G Della Valle, M Malig, M Rocchi, E E Eichler, and M Ventura. New insights into centromere organization and evolution from the white-cheeked gibbon and marmoset. *Molecular biology and evolution*, 26 :1889–1900, August 2009.

[6] H F Willard. Evolution of alpha satellite. *Current opinion in genetics & development*, 1 :509–514, December 1991.

[7] D M Kurnit and J J Maio. Variable satellite dna's in the african green monkey cercopithecus aethiops. *Chromosoma*, 45 :387–400, May 1974.

[8] C Lee, R Wevrick, R B Fisher, M A Ferguson-Smith, and C C Lin. Human centromeric dnas. *Human genetics*, 100 :291–304, September 1997.

[9] I Alexandrov, A Kazakov, I Tumeneva, V Shepelev, and Y Yurov. Alpha-satellite dna of primates : old and new families. *Chromosoma*, 110 :253–266, August 2001.

[10] Valery A Shepelev, Alexander A Alexandrov, Yuri B Yurov, and Ivan A Alexandrov. The evolutionary origin of man can be traced in the layers of defunct ancestral alpha satellites flanking the active centromeres of human chromosomes. *PLoS genetics*, 5 :e1000641, September 2009.

[11] L Y Romanova, G V Deriagin, T D Mashkova, I G Tumeneva, A R Mushegian, L L Kisselev, and I A Alexandrov. Evidence for selection in evolution of alpha satellite dna : the central role of cenp-b/pj alpha binding region. *Journal of molecular biology*, 261 :334–340, August 1996.

[12] Mark S Springer, Robert W Meredith, John Gatesy, Christopher A Emerling, Jong Park, Daniel L Rabosky, Tanja Stadler, Cynthia Steiner, Oliver A Ryder, Jan E Janečka, Colleen A Fisher, and William J Murphy. Macroevolutionary dynamics and historical biogeography of primate diversification inferred from a species supermatrix. *PloS one*, 7 :e49521, 2012.

[13] Andrej Benjak, Claudia Sala, and Ruben C Hartkoorn. Whole-genome sequencing for comparative genomics and de novo genome assembly. *Methods in molecular biology (Clifton, N.J.)*, 1285 :1–16, 2015.

[14] Robert C. Edgar. Muscle : multiple sequence alignment with high accuracy and high throughput, 2004.

[15] Peter Rice, Ian Longden, and Alan Bleasby. Emboss : the european molecular biology open software suite, 2000.

Résumé

Chez les primates, des séquences centromériques répétées en tandem sont appelées ADN α-satellite. Un monomère α-satellite fait 171 pb de long. Elles ont un taux d'identité de 60% à 100%. Ces monomères peuvent être regroupés en familles selon la similarité. Pour étudier ces familles, des méthodes utilisant la phylogénie existent mais elles ne permettent pas de classer un grand nombre de séquences. De plus les méthodes ne sont pas objectives et ne permettent pas de comparer les espèces entre elles. Pour pallier ce problème,l'équipe ARChE a développé une méthode de classification permettant de traiter des centaines de milliers de séquences (2017). Mon sujet consiste à appliquer cette méthode aux jeux de données déjà publiés pour évaluer la méthode. Dans un deuxième temps, cette méthode est appliquée à deux autres primates dans le but de caractériser les familles d'espèces proches. Les mécanismes d'évolution sont déduits à partir d'une comparaison inter-espèce révélant les différences et les familles communes.

Abstract

Centromeric repeated sequences in Primates are named α-satellite DNA. An α-satellite's length is about 171 pb. Its identity rate is between 60% and 100%. Monomers can be gathered into families according to their similarity. To study these families, methods using phylogeny are used but a large number of sequences cannot be processed. Moreover these methods are not objective and do not allow inter-species comparison.To overcome this problem, the team ARChE developped a classification method which processes hundred of thousands of sequences (2017). The subject of my internship is to apply this method to published datasets and assess this method. Then this method is applied to other Primates in order to caracterize families in close species. Mechanism of evolution are deducted from inter-specices comparison, revealing common and different families.